DOCUMENT RESUME

ED 283 856 TM 870 378

AUTHOR Wolfe, Mary L.
TITLE A Bayesian Approach to Predicting Academic Performance.
PUB DATE [Jan 87]
NOTE 14p.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Academic Failure; *Bayesian Statistics; *Discriminant Analysis; Expectancy Tables; *Grade Prediction; Higher Education; Multiple Choice Tests; *Predictive Measurement; Statistical Studies; *Student Evaluation
IDENTIFIERS Nursing Students

ABSTRACT

A total of 149 students enrolled in an undergraduate nursing research methods course participated in a study comparing three strategies for using formative evaluation (test feedback throughout a course) to predict students at risk of failure at summative evaluation (the final examination). Students took 12 weekly multiple-choice quizzes, which were graded and returned for self-study, and a final 60-item multiple-choice exam. Three 4-week quiz subtotals were the discriminating variables used to predict membership in three final-exam score categories: Group 1 (poor); Group 2 (fair); Group 3 (good). Separate discriminant analyses tested three patterns of assigning prior probabilities of group membership: (1) equal (each .333); (2) proportional to actual numbers of students in each group; (3) weighted by setting cost of misclassifying poor students as three times more serious than cost of misclassifying fair or good students. A significant discriminant function emerged, and confirming previous results, effect size (a standardized measure of the discrepancy between performance and the overall mean) for poor students decreased over time, showing that they were "closing the gap." Assigning probabilities proportional to cases gave best overall classification accuracy (53.02%), but Bayesian weighted adjustment best predicted students at risk of failure (82.1% correctly classified) while sacrificing some overall predictive power (42.95% correct). (LPG)

A Bayesian Approach to Predicting Academic Performance

Mary L. Wolfe, Ph.D

School of Nursing

University of Maryland

Baltimore

Frequent testing with instructor-made tests has become a common practice in many college and universities, especially in courses such as research design or statistics where the acquisition of a hierarchy of skills is required. In courses where short quizzes are given during each class period, a significant proportion of class time is given over to such testing during the semester. It is important that this practice be evaluated in terms of its utility in promoting learning, improving instruction and identifying learning problems which may require intervention.

The use of frequent quizzes to monitor progress is an example of what Bloom, Hastings and Madaus (1971) have called formative evaluation. Formative evaluation entails the collection of relevant data to provide guidance for the learner, and indicate the need for modifications in teaching strategies. Used properly, it can improve the instructor's ability to meet individual needs. The major goal of formative evaluation is to provide feedback during the learning process on errors and misconceptions, rate of progress, and achievement relative to an acceptable level of competence. Summative evaluation, in contrast, provides a general assessment of student achievement over an entire course or large unit and is usually the major determinant of course grades.

According to Bloom, one potential use of formative evaluation

data is in predicting the outcome of summative evaluation. Since there is usually considerable overlap between the two kinds of assessment in terms of content, behaviors and testing procedures, the two kinds of test results are likely to be highly correlated. Thus, it may be possible to predict performance on summative tests in advance, and to alter the prediction for the better. Empirical evidence suggests that students use data from formative evaluation to modify their study habits and improve their performance over time. Wolfe (1981) in a study of students in an undergraduate nursing research course, found that prediction of midterm examination scores from pre-midterm quizzes was considerably more accurate than the prediction of final examination scores from pre-final exam quizzes. She concluded that by the time half a term had passed, students had learned to use the results of weekly quizzes to change the prediction as Bloom suggested. In another study with the same population (Wolfe, 1985) weekly quiz scores were summed over each of four three-week periods. Final exam scores were dichotomized as satisfactory (A or B) or unsatisfactory (C or below). Discriminant analyses revealed that 71% of the final examination scores were correctly classified as satisfactory or unsatisfactory on the basis of quiz subtotal scores. Examination of group differences on the individual discriminating variables showed that mean scores for the "satisfactory" group were higher on the first three quiz subtotals than those of the "unsatisfactory" group. However, during the last quarter term this difference was reversed, with the "unsatisfactory" group

achieving a slighly higher mean quiz subtotal. This finding suggested that students having more global difficulty with course material, as reflected in lower final examination scores, may have tried somewhat harder to modify their study habits and improve their standing toward the end of the term, compared to their relatively secure classmates.

In a third study (Wolfe, 1986) the extent to which grades on a comprehensive final examination could be classified as good (A or B), fair (C) or poor (D or F) on the basis of weekly quiz scores was determined. Students in an undergraduate nursing research course were given 12 short weekly quizzes and a comprehensive final examination. Three discriminant analyses were performed with recoded final examination scores as the grouping variable. Discriminating variables for the analyses consisted of the four quiz scores from the first, middle and last third of the semester, respectively. Although quiz scores discriminated significantly between the groups for each time period ($p<.05$), the percentage of cases correctly classified decreased over time. The fact that final examination grades became less predictable over time further supported Bloom's conjecture that students may indeed modify their study habits and change the forecast.

In using discriminant analysis to predict group membership on the basis of a set of measurements, an individual is assigned to that group for which he or she has the highest posterior probability of membership. Most computer programs (e.g., SPSS-X, 1986) offer several options for determining the so-called prior probability of group membership. The prior probability of a given

population is the probability that an individual selected at random actually comes from that population. For instance, in a three-group discriminant analysis in which the groups are equal in size, a straightforward assumption would be that a person selected at random has a prior probability of one-third of being classified into any one of the groups. That is, without knowing any of the individual's characteristics, we are equally likely to classify him or her as belonging to group 1, 2 or 3. However, a Bayesian adjustment of this prior probability may be advisable if the group sizes differ widely or if the costs of misclassification into certain groups are considered very high. For instance, in the case of students who are at risk of academic failure or poor performance, the cost of failing to identify them early may be regarded as several times greater than the cost of misclassifying students whose performance is satisfactory.

The purpose of the present study was to compare the effects of three different procedures for specifying the prior probabilities of group membership on improving the ability to correctly classify students at risk of failure in an undergraduate research methods course.

METHODS

Sample. One hundred forty nine students in five sections of an undergraduate course in nursing research methods participated in the study. All sections were taught by the investigator during 1984 and 1985.

Procedure. For each section, 12 multiple-choice quizzes with five to ten items were given, one each week after the first week of class. Each quiz covered content which had been presented and reviewed during the previous class session. Students exchanged completed quizzes with their neighbors for grading, and the correct answers were read in class by the instructor. The quizzes were returned to the examinees, and each question was discussed in as much detail as needed. Following review, quizzes were collected and grades recorded by the instructor. The following week, quizzes were returned to students with the suggestion that they keep them to aid in reviewing for the final examination. The course content was the same in each section, with the first half term devoted to descriptive and correlational statistics, measurement and research design, while the second half dealt with statistical inference. On the last day of class a comprehensive 60-item multiple choice examination was given. Students were invited to contact the instructor to arrange for individual conferences regarding their performance.

RESULTS

For the purpose of statistical analysis, quiz scores were summed over each of three four-week periods, in order to enhance predictor reliability and ensure a favorable ratio of subjects to variables. Final examination scores were re-coded as follows: Group 1, poor (41 or below); Group 2, fair (42-47); Group 3, good (48-60).

Three stepwise discriminant analyses were performed on the data, with the recoded final examination scores as the dependent measure and the three four-week quiz subtotals as the discriminating variables. For the first analysis the prior probabilities were all assumed to equal 0.3333. For the second analysis, the prior probabilities were specified as the proportions of cases in each group: for Group 1, 0.2617; for Group 2, 0.4295; for Group 3, 0.3087. For the third analysis, a procedure suggested by Afifi and Clark (1984) was followed. The investigator considered it three times as serious to misclassify a poor student as it was to misclassify a fair or good student. Thus, the proportions of cases in groups 1, 2 and 3 were multiplied by 3, 1 and 1, respectively:

For Group 1: adjusted p_1 = .2617 X 3 = .7851

For Group 2: adjusted p_2 = .4295 X 1 = .4295

For Group 3: adjusted p_3 = .3087 X 1 = .3087

Since the prior probabilities must sum to 1, the probabilities computed above were further adjusted by dividing each one by .7851 + .4295 + .3087 = 1.5233. Thus, the final values for the prior probabilities were:

For Group 1: q_1 = .515

For Group 2: q_2 = .282

For Group 3: q_3 = .203

The results of the three analyses are shown in Table 1.

Table 1 about here

One significant discriminant function was found (χ^2 = 35.297, df = 4, p = .0000).

DISCUSSION

The fact that quiz subtotals discriminated significantly among students classified as poor, fair or good on the basis of final examination performance is in accord with results obtained earlier by the same investigator. Examination of univariate F-ratios showed that the groups differed significantly during all three periods ($p<.05$). The F-ratios were considerably larger during the first two periods (F = 11.71 and F = 13.13, respectively) than during the last period (F = 3.82), suggesting that as in the earlier studies, the weaker students may have attempted to close the gap between their performance and that of their stronger peers. This observation was validated by computing "effect sizes" for each group for each time period (See Table 2). Effect size was computed by subtracting the grand mean

---

Table 2 about here

---

for each time period from the group mean, and dividing this difference by the total standard deviation for all 149 cases. For the "poor" group, the effect size - a standardized measure of the discrepancy between their performance and the overall mean - showed a small but consistent decrease from the beginning to the end of the term.

Of major interest is the effect of adjusting the prior probabilities on the percentage of students at risk of failure correctly classified. Table 1 shows that, although the largest overall percentage of cases correctly classified (53.02%) was obtained when the prior probabilities were made proportional to group sizes, the "poor" group, because of its relatively small size, was assigned the smallest prior probability. As a result, only 38.5% of the students in that group were correctly classified - nearly two-thirds would have been incorrectly identified as "fair" or "good" performers. In contrast, when the cost of misclassification of poor performers was taken into account and the prior probabilities adjusted accordingly, 82.1% of the students in this group were correctly classified. However, the overall percentage of cases correctly classified was only 42.95%.

Clearly, there is a tradeoff involved in the decision to weight prior probabilities according to the perceived cost of misclassification. The method selected must be guided by the purpose of the statistical analysis as well as the personal philosophy of the investigator. One must weigh the potential harm done to a good student who is mistakenly informed that he or she in academic jeopardy against the possibly greater damage which would occur if a student at risk of failure is not identified early enough for effective intervention.

REFERENCES

Afifi, A.A. and Clarke, V. Computer-aided multivariate analysis. Belmont, CA: Lifetime Learning Publications, 1984.

Bloom, B.S., Hastings, J.T. and Maduas, G.F. Handbook on formative and summative evaluation. New York: McGraw-Hill, 1971.

SPSS-X. User's guide (2nd edition). New York: McGraw-Hill, 1986.

Wolfe, M.L. Forecasting summative evaluation from formative evaluation: a double cross-validation study. Psychological Reports, 1981, 49, 843-848.

Wolfe, M.L. Predicting final examination performance in an undergraduate nursing research course. Poster session presented at the Annual Meeting of the Society for Research in Nursing Education, January 1985, San Francisco, CA.

Wolfe, M.L. Predicting summative evaluation from formative evaluation in a baccalaureate nursing research course. Poster session presented at the Annual Meeting of the Society for Research in Nursing Education, January, 1986, San Francisco, CA.

Table 1. Classification results for analyses 1, 2 and 3

Analysis 1 (priors equal)

Actual group	No. of cases	Predicted group membership		
		1	2	3
Group 1	39	25	5	9
		64.1%	12.8%	23.1%
Group 2	64	21	18	25
		32.8%	28.1%	39.1%
Group 3	46	3	13	30
		6.5%	28.3%	65.2%

Percent of grouped cases correctly classified: 48.99%

Analysis 2 (priors proportional to group size)

Actual group	No. of cases	Predicted group membership		
		1	2	3
Group 1	39	15	20	4
		38.5%	51.3%	10.3%
Group 2	64	11	40	13
		17.2%	62.5%	20.3%
Group 3	46	2	20	24
		4.3%	43.5%	52.2%

Percent of grouped cases correctly classified: 53.02%

Table 1 (continued)

Analysis 3 (priors weighted by cost of misclassification)

Actual group	No. of cases	Predicted group membership		
		1	2	3
Group 1	39	32	3	4
		82.1%	7.7%	10.3%
Group 2	64	42	9	13
		65.6%	14.1%	20.3%
Group 3	46	18	5	23
		39.1%	10.9%	50.0%

Percent of grouped cases correctly classified: 42.95%

Table 2. Effect sizes for the three groups at each time period.

	1st period	2nd period	3rd period
Group 1	-0.553	-0.494	-0.332
Group 2	0.028	-0.076	0.018
Group 3	0.430	0.525	0.257